Computer-aided content analysis and "soft data" in historical social research: an attempt to find a pragmatic solution

Breyer, Gerald; Finzsch, Norbert; Schaefer, Jochen; Straeter, Johannes; Wengler, Frank; Wisniewski, Birgit

Historical Social Research, Vol. 15 — 1990 — No. 3, 206-213

COMPUTER-SECTION

Computer-Aided Content Analysis and »Soft Data« in Historical Social Research

An Attempt to Find a Pragmatic Solution

*Gerald Breyer, Norbert Finzsch, Jochen Schaefer, Johannes Straeter, Frank Wengler and Birgit Wisniewski(l)**

As Ekkehard Mochmann pointed out more than a decade ago, a content or document analysis is a computer-aided procedure which has established a sufficiently long tradition to be recognized as a standard instrument in social research. (2) In concordance with Mochmann, we define content analysis for the purpose of this paper as a procedure comprising the systematic description, reduction and inspection of communication in an analytic framework of research concepts. (3) It is obvious that content analyses can be conducted by using computers. One may even go further and say that before the application of machines for that purpose, large-scale inquiries remained the exception to the rule. Sola Pool writes about this period of scientific history, when there were no machines to do the hard work of retrieval and counting:»1 stopped doing content analysis before Phil Stone had developed the General Inquirer, because it was too hard. The amount of work involved for the product was enormous.« (4)

So computers have turned out to be great time savers, but in addition to that, they can enhance the precision of the analysis, make the process reproducible, and help to avoid mistakes as a result of fatigue or stress. (5) Different programs have been developed in order to achieve these improved results, most of which had been conceptualized in the second half of the sixties, at a time when nobody thought of the computer revolution. Most prominent are General Inquirer by Philip James Stone, the pioneer in this field, EVA, Textpack and others. (6)

* Address all communications to: Norbert Finzsch, German Historical Institute, 1607 New Hampshire Ave., N. W., Washington, D. C. 20009, USA.

During the past ten past years, the discussion on computer aided content analyses focused around the methods of creating a machine-readable corpus of data which would yield the appropriate material for the formulation of categories for the content analysis itself. The problem was to find the optimal combination of associationally rich words which »contained« the information the researcher was looking for. Thus a new text was created out of the original one both by removing all words which had only a grammatical function and by constructing a meta-text in which all the redundancies - so typical of natural languages - have been left out. In this meta-text only the meaningful components have been preserved before the content analysis itself can begin. The researcher must determine what a meaningful utterance actually is. One of the motives for such an approach was the need to reduce the length of a text so that it would be easier to handle. Long texts are more difficult to analyze, and it is very expensive to transform them into machine-readable sets of data. With the development of optical character reading scanners (OCR), however, it has become far more likely that in the future one will be able to produce content analyses even with very long texts without first having them to tranform into aesthetically and linguistically poorer meta-texts.

The other reason which was thought to be appropriate for reducing texts to a more abstract level, was the tacit and now obsolete assumption that function words and lemmata do not have a significance for the semantic structure of the text. This is a conviction that is not even shared by a majority of linguists today, not to speak of literary critics or communication theorists. (7) Although it would be much easier for social scientists if they could strip a text of its »noise« and reduce it to its pure »signals«, one has to concede that this is not the way human communication works through texts, whether in literature or in every day life. The communication triangle model which referred to technical processes (transmitter - code - receiver) was a fallacy.

At present, any ideas of automatic speech understanding by computers have to be dismissed, because processors able to »understand« human actions would have to include the knowledge of contextuality and pragmatics. One of the prerequisites for automatic speech understanding would be a machine-readable dictionary of all the words expected in a text, together with their flexations, suffixes and prefixes (which is in itself a eurocentric attitude, because several non-Indoeuropean languages do not have these), before the computer could begin to scan through the texts, sorting sentences (another controversal concept - in terms of linguistics) and »meaning«. It is possible, even today, to develop a machine with a certain set of »expectations«, with a collection of variations on what a human would say in a given situation/context/action. We do not doubt the possibility of setting up a voice-controlled baggage sorter at Chicago's O'Hare Airport. (8) We

doubt, rather, that even the most sophisticated word cruncher machine will be able to analyze sources from the sixteenth century, without a »human interfaces who had previously read and understood the sources. The definition given in the opening paragraph of this paper must therefore be modified. Content analysis is not understood anymore as objective, systematic, and quantitative decription of the manifest content of communication, but rather as a research technique that allows one to draw inferences through systematic and objective identification of clearly defined characteristics of a text. (9) Opinions about the manifest content of a given communication may differ to a great extent. For the purpose of this paper it is sufficient to point out that communication implies the changing of meanings, state of information, values and attitudes as well as behavior of those involved in it. (10) Such definition excludes the possibility of »automatic« text analyses. (11)

Discussions about content analyses have developed in two directions since the 1980s. First, the knowledge about communication available for social scientists is far more sophisticated and complex than it was a decade ago. Second, microcomputers are far more advanced and have become an everyday tool for most researchers in a way unimagined in 1980. In a certain way this has led to a concentration upon the application of content analysis to IBM-compatible computers. Programs like SELECT, hotly debated in the 1970s, almost seem to be technological dinosaurs nowadays, when compared to modern user-friendly programs that fit onto a single diskette and can be handled by first-year students without first having to attend a class in advanced programming.

On a different level, these contingencies have led the scientific community to assume that there are two dominant forms of content analysis, i.e. automatic content analysis and analysis with predetermined descriptors. The latter method is by far the most traditional and was already developed in social science research projects in the 1930s. In this paper, we refer to this latter form exclusively, since automatic procedures are out of reach, as pointed out before. For this traditional approach, one does not necessarily need a computer, but it helps a great deal when one can work with a little help from this friend. All one must do is to construct a predefined category scheme with intentionally and explicitly-circumscribed research variables. Then the researcher must scan the material with this set of variables in mind, provided that a pretest has proved the variables to be valid and adequate. (12) It is a common practice among content analysts to produce non-stable and thus non-reproducible results because their focus of perception changes over time, as they come to understand similar or identical passages in the text in a different way. That, of course, has tremendous implications for the de facto definition of variables with which one wants to analyze the text. Even if a sufficiently large sample of

the text is taken for a pretest, there may be distortions of the original scope of a variable through hermeneutical processes, tending to sidestep the initially clearly defined system of variables. For reasons of fairness we refer here to our own work only. (13)

The traditional method, cited above, is rather impracticable, because the text under scrutiny must be indexed according to predetermined keywords. Let us consider an example. Norbert Finzsch has conducted a study of almost 1.700 letters written on behalf of poor persons to the overseers of the poor (bureau de bienfaisance) in the city of Cologne at the beginning of the nineteenth century. (14) One question raised was whether there was a modern understanding of poverty as opposed to the the traditional one, influenced by the Christian doctrines of Thomas Aquinas and others. Modern in this case was equated with an understanding of the social causes of poverty, i.e. lack of work, illness of the breadwinner, or old age which did not allow the person to work anymore. A combination of these reasons was possible. Thus, the researcher must decide whether a text made reference to the causes of poverty. The variable POVERTY would thus receive a postive or negative marking. In a second step, the researcher had to decide whether the reason given for poverty was to be found in the sphere of WORK, ILLNESS, AGE etc. Thus one could develop a matrix of content in the following form:

POVERTY	YES NO
if answer is yes:	
WORK	YES NO
ILLNESS	YES NO
AGE	YES NO

or in a more abstract form:

POVERTY	1
WORK	1
ILLNESS	0
AGE	1

where »1« means »yes« and »0« means »no«, indicating in this case that the text asserts reasons for poverty, lying in the area of work (joblessness) and age (person being too old to get a new job). After the source material had been scanned and the content matrix of each document had been keypunched, a statistical analysis of the whole collection of documents was possible.

As can be understood by this example, the categorization of the text through keywords resists later redefinitions of indices. (15) Therefore a thorough investigation of the text, its »meaning« within the context of the study under way, and an understanding of the historical episteme are ne-

cessary, before a system of categories for the final analysis can be developed. A striking example is the changing meaning of the term »Handarbeit« (handwork) in the eighteenth century. During the 1750s, that term meant meant hard, physical labor; by the end of the century, it acquired a second connotation, that of needlework. A second example of the need for a hermeneutical grasp of the text is the connotation of the word »Mensch« (man/human being) which in some local dialects in the Rhineland referred exclusively to females. It is obvious that in research pertaining to the social roles of individuals it is extremely important to know whether a word refers to a male or a female.

A pragmatic approach to the problems investigated here is to apply programs facilitating the laborious task of content analysis by turning most steps over to the machine, while the basic thinking still must be done by the human. By the term »pragmatic« we mean a) a practical, easy-to-use, and efficient way; and b) a method which takes into account the pragmatic aspects of the text - in the linguistic sense. (Here one finds another example for the problems posed by using tropes in a text - this one would be a hard nut to crack for automatic content analyses. (16) We did some practical research on the applicability of one of the products on the market by using it in a study previously conducted in the traditional way. Norbert Finzsch had produced a set of data for statistical analysis by going through the original sources and transferring the data first onto coding lists and then onto data carriers which were then computed on a mainframe computer with SPSS + . The most important procedure of the study, the coding of natural language into semantic matrices of the meta-text, in this case was done as »handwork«. The application of the computer only began after the data were already assembled. The problem with such an approach is, that it is hard to find valid and adequate sets of categories on the first try - unless one draws a representative sample and conducts a very time-consuming pretest. (17)

Two years after this study had been completed, we took the same source material again and used a concordance analyzer called CONCORD, in order to »know« the text before we developed a set of criteria or categories with which to analyze it. The sources, a random sample of 330 letters drawn from the complete set of 1.700 analyzed by Finzsch in 1987, were transcribed and processed by CONCORD as a machine-readable file on a IBM-AT compatible computer. It produced about 600 pages of word lists consisting of every single word that occurred in the text in its phrasal context in alphabetical order. Thus we could begin to isolate those expressions which seemed to carry »meaning« in regard to the questions at hand. Close reading of those semantic carriers proved that a considerable number of those words must have changed their meanings dramatically over the past 200 years. Fortunately CONCORD has a very fast and precise

word-in-context retrieval algorithm, so that it was possible to compare different uses of one word in their respective contexts. By looking at the contexts, we were able to assign the old meanings to these words, and only after this procedure had been terminated, could we start to assemble semantic matrices and categories for the statistical evaluation of the text. We could bundle sets of words into one semantic cluster only because we had a total overview of all the words that appeared in the text. Those clusters were then combined into variables with either nominal or real scalings. The coding itself was then done by hand again, and the data were processed by SPSS/PC+ as in the first case, only this time on a personal computer.

The results were astounding. Checking the results of both studies with the original texts clearly demonstrated the higher efficiency of the concordance-content analysis. Although only a sample had been chosen from the whole corpus, the statistical accuracy, validity, and adequacy of the second study was much higher than in Finzsch's earlier study, in which the complete and unbroken set of sources had been used. (18) By allowing a step-by-step reduction of the meaning which was controlled by a complete lexematic compilation of all words through the concordance and a total operationalization of concepts in the texts, we were able to make the content analysis reproducable and its results thus were stabilized, while the traditional content analysis tended to produce a rather broad variation of meaning assignments. By criticizing the two extreme approaches to content analyses, the inefficient fault-ridden traditional one and the epistemologically problematic automatic content analysis, we came up with a kind of middle-of-the-road solution, using the old device of a (automatic) concordance in combination with manual coding and statistics put together by conventional software.

Notes

1 The research leading to this article was done by my Bochum students, which are consequently listed as collective coauthors. I have to claim responsibility for the actual wording of this article. My colleague Ken Ledford was a great help in editing my awkward English sentences. For correspondence please write to Norbert Finzsch, German Historical Institute, 1607 New Hampshire Avenue, N.W., Washington D.C. 20009, U.S.A.

2 Ekkehard Mochmann, Methode und Techniken automatisierter Inhaltsanalyse, in: ibid. (ed.), Computerstrategien für die Kommunikationsanalyse, New York, Frankfurt 1980 (Beiträge zur empirischen Sozialforschung), p. 13. For a good introduction to the theoretical

problems involved in content analysis in English see Klaus Krippendorff, Information Theory: Structural Models for Qualitative Data, Beverly Hills 1986. More user-oriented is ibid., Content analysis: An Introductiuon to Its Methodology, Beverly Hills 1980.

3 Ekkehard Mochmann, Computer Aided Content Analysis of Historical and Process-Produced Data: Methodological and Technical Aspects, in: Jerome M. Clubb; Erwin K. Scheuch, Historical Social Research: The Use of Historical and Process-Produced Data, Stuttgart 1980 (Historisch-sozialwissenschaftliche Studien, vol. 6), pp. 235-243, p. 236.

4 Sola Pool, Bridging the Gap between Content Analysis and Survey Research, in: Mochmann (ed.), Computerstrategien, p. 245-248.

5 Holger Rust, Inhaltsanalyse: Die Praxis der indirekten Interaktionsforschung in Psychologie und Psychotherapie, München, Wien, Baltimore, 1983, p. 121-141.

6 Philip James Stone et al., The General Inquirer: A Computer Approach to Content Analysis, Cambridge, Mass., 1966.

7 For a general criticism of the methods of content analysis by linguists see Ingunde Fühlau, Die Sprachlosigkeit der Inhaltsanalyse: Linguistische Bemerkungen zu einer sozialwissenschaftlichen Methode, Tübingen 1982. See also Peter Möhler, Computergestützte Inhaltsanalyse zwischen Algorithmen und Mythen, in: Sprache und Datenverarbeitung 9, 1985, pp. 11-14. It is impossible to quote all the essential literature on the problem. We refer to Deborah L. Dennis, »Word Crunching«: An Annotated Bibliography on Computers and Qualitative Data Analysis, in: Qualitative Sociology 7, 1-2, 1984, pp. 148-156. More recent, but available only as »grey« literature is Peter Möhler; Katja Frehsen; Ute Hauck, Computergestützte Inhaltsanalyse: Grundzüge und Auswahlbibliographie zu neueren Anwendungen, in: ZUMA-Arbeitsbericht, vol. 9, Mannheim 1989.

8 Georgette Silva, On Automatic Speech-Understanding, in: Computers and the Humanities, vol. 9, 1975, pp. 237-244.

9 B. Berelson, Content Analysis in Communication Research, Glencoe, Illinois 1952, p. 18. Mochmann (ed.) Computerstrategien, p. 13.

10 Mochmann (ed.), Computerstrategien, p. 9.

11 Reductionist theoretical models of content analysis have been widely criticized. See Hans-Dieter Klingmann (ed.), Computerunterstützte Inhaltsanalyse in der empirischen Sozialforschung (Zuma Monographien Sozialwissenschaftliche Methoden, vol 4.), Frankfurt 1984.

12 The problem of adequacy and validity of variables for content analyses has been thoroughly and allembracingly discussed by Klaus Krippendorff, so that we do not want to delve deeper here. Klaus Krippendorff, Validity in Content Analysis, in: Mochmann (ed.), Computerstrategien, p. 69-112, here p. 80-87.

13 Heike Bernhardt, Norbert Finzsch et al., Erfahrungsbericht der Arbeitsgruppe »Armutsforschung«, in: HSR 13,3, 1988, pp. 163-171. Norbert Finzsch, Obrigkeit und Unterschichten: Beiträge zur Geschichte rheinischer Unterschichten gegen Ende des 18. und zu Beginn des 19. Jahrhunderts, Stuttgart 1990, in which a conventional content analysis was conducted with almost 1.700 letters written to the overseers of the poor in Cologne between 1799 and 1802.

14 Finzsch, Obrigkeit und Unterschichten, as in footnote 13..

15 Margarete Höllbacher, EDV-Programme für die sozialwissenschaftliche Textanalyse: Versuch zur Erstellung eines Anforderungskatalogs, in: Heinz Jürgen Kaiser; Hans-Jürgen Seel (eds.), Sozialwissenschaft als Dialog: Die methodischen Prinzipien der Beratungsforschung, Weinheim, Basel 1981, pp. 241-260. Höllbacher also discusses the methodological problems of automatic analyses.

16 It is a question to be solved in how far computers will be able to reproduce syntactical structures of human parole. For a discussion of the state of art see JanJ. van Cuilenburg, J. Kleinnijenhuis, J.A. den Ridder, Artificial Intelligence and Content Analysis: Problems of and Strategies for Computer Text Analysis, in: Quality and Quantity 22,1, 1988, pp. 65-97.

17 Werner Früh, Konventionelle und maschinelle Inhaltsanalyse im Vergleich: Zur Evaluierung computergestützter Bewertungsanalysen, in: Hans-Dieter Klingmann (ed.), Computergestützte Inhaltsanalyse, p. 35-53.

18 For a short introduction to CONCORD see Norbert Finzsch, CONCORD - A Program for Concordance Analyses on a Personal Computer, in: Historical Social Research 14,4, 1989, p. 156-158.